1-2013

Utility of Potential Misdiagnoses in Predicting Foodborne Outbreaks

Lucia LUCIA
Singapore Management University

Artur DUBRAWSKI
Carnegie Mellon University

Lujie CHEN
Carnegie Mellon University

Utility of Potential Misdiagnoses in Predicting Foodborne Outbreaks

Lucia Lucia[1*], Artur Dubrawski[2], Lujie Chen[2]

[1] Singapore Management University, Singapore; [2] Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA.

Objective

To investigate utility of using inpatient and emergency room diagnoses to detect outbreaks of Salmonellosis in humans. To quantify the impact of including in the analysis cases diagnosed with conditions that may have physiological appearance similar to Salmonellosis.

Introduction

Reliable detection and accurate scoping of outbreaks of foodborne illness are the keys to effective mitigation of their **impacts**. However, relatively small number of persons affected and underreporting, challenge the reliability of surveillance models. In this work, we correlate a record of identified outbreaks and sporadic cases of Salmonellosis in humans retained in PulseNet [1], and diagnosis codes in hospital claims collected in California from 2006 to 2010. We hypothesize that the data support and reliability of detection could be improved by including cases in which Salmonella infection may be confused [2].

Methods

We join the data in a table indexed with dates and locations, containing counts of inpatient and ED patients diagnosed with Salmonellosis and related diseases, also counts of cases involved in outbreaks, aggregated by day (the admission date or the isolation date) and location (the county of hospital locations or the county where the outbreaks occurred). 9.5% of the 66,845 rows in the table involve sporadic cases and identified clusters.

To quantify predictive utility of potential misdiagnoses, Zero-inflated Poisson regression (ZIP) model [3] is trained to predict the number of cases in epidemiological data. Among Salmonellosis (counts in inpatient and ED) and 12 potential misdiagnoses, the best combination of input features is found by exhaustive search to minimize 10 fold cross validation ZIP prediction error. The chosen model is then trained using thusly selected features using all data. Similarly, we train a Random Forest (RF) binary classifier [4] that also includes spatio-temporal predictors (county and month) to discount seasonality and spatial propensity of outbreaks.

Results

We found that 8 diagnoses related to Salmonellosis have non-trivial impact on outbreak predictability (only Celiac is insignificant with p-value>0.05). Their contributory effect is indicated by positive coefficients of ZIP count model and negative coefficients of ZIP zero model, as shown in the table.

Including counts of these diagnoses improves predictability of the occurrence of outbreaks vs. using Salmonellosis diagnoses only. The AUC score of the RF model increases from 57% to 87%. Adding spatio-temporal factors improves the predictability to 91% AUC. The model discovers 71% of actual outbreak cases at 7% false positive rate (FPr) and correctly recalls 4.5 as many outbreak cases at 1% FPr as when using Salmonellosis diagnoses only.

We found 37% of the predictions can be made 1 to 7 days earlier than the recorded isolation date, increasing precision to 89%. This suggests a potential early warning utility. It is also possible to spot outbreaks not revealed in Pulsenet. For instance, 22 of 35 outbreak predictions in Yolo County are not in PulseNet; 60% of these 22 have at least 40% of nearby counties showing positive predictions or actual cases in Pulsenet in the same periods of time.

ZIP Model	Intercept	Salmonellosis emergency	Salmonellosis Inpatient	Shigellosis
Count	-0.477639	0.3227	0.2897	0.1430
Zero	3.47364	-1.1529	-1.0431	-0.0944

ZIP Model	Crohns	Diabetic gastroparesis	Colon cancer	Diarrhea
Count	0.0463	0.0424	0.0348	0.0248
Zero	-0.0445	-0.1459	-0.3631	-0.2986

ZIP Model	Celiac	Ulcerative colitis	Viral gastroenteritis
Count	0.0235	0.0215	-0.0368
Zero	-0.4505	-0.2772	0.0219

Conclusions

We empirically found informative correlation between the counts of hospital patients diagnosed with diseases that may have physiological appearance similar to Salmonellosis, and epidemiologically recorded cases of Salmonellosis. This suggests that tracking these diseases could support accuracy of foodborne illness surveillance. Further study is yet required to verify the actual extent of clinical misdiagnosing, and if there are other factors explaining the apparent correlation.

Keywords

Foodborne outbreaks; misdiagnosis; predictive analytics.

Acknowledgements

This work is supported by the National Science Foundation (awards 0911032, 1320347) and the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority.

References

1. PulseNet. http://www.cdc.gov/pulsenet/about/index.html
2. Rightdiagnosis. http://www.rightdiagnosis.com
3. Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. 1992.34(1): 1-14.
4. Breiman L. Random Forests, *Machine Learning*. 2001.45(1).

*Lucia Lucia
E-mail: Lucia.2009@phdis.smu.edu.sg